1.2: Classifying and Storing Data

- The collection of data is called a **data set** or a **sample**. The **population** refers to the set or group that contains everything relevant to the data.
- When we collect data, the characteristics of that data (e.g. gender, weight, temperature) are called **variables**.
- Variables can be categorized into two groups:
  - Numerical variables — Quantitative charateristics: length, temperature, age
  - Categorical variables — Qualitative characteristic: Color, area code, size (S, M, L), shape

Example. The following table contains data crash-test dummy studies.

- How many variables does this table have? 5
- How many observations does this table have? 9
- For each variable, identify whether it is numerical or categorical:

Cat.	Cat.	Num. (maybe Cat.)	Num.	Num.
Make	Model	Doors	Weight	Head Injury
Acura	Integra	2	2350	599
Chevrolet	Camaro	2	3070	733
Chevrolet	S-10 Blazer 4X4	2	3518	834
Ford	Escort	2	2280	551
Ford	Taurus	4	2390	480
Hyundai	Excel	4	2200	757
Mazda	626	4	2590	846
Volkswagen	Passat	4	2990	1182
Toyota	Tercel	4	2120	1138

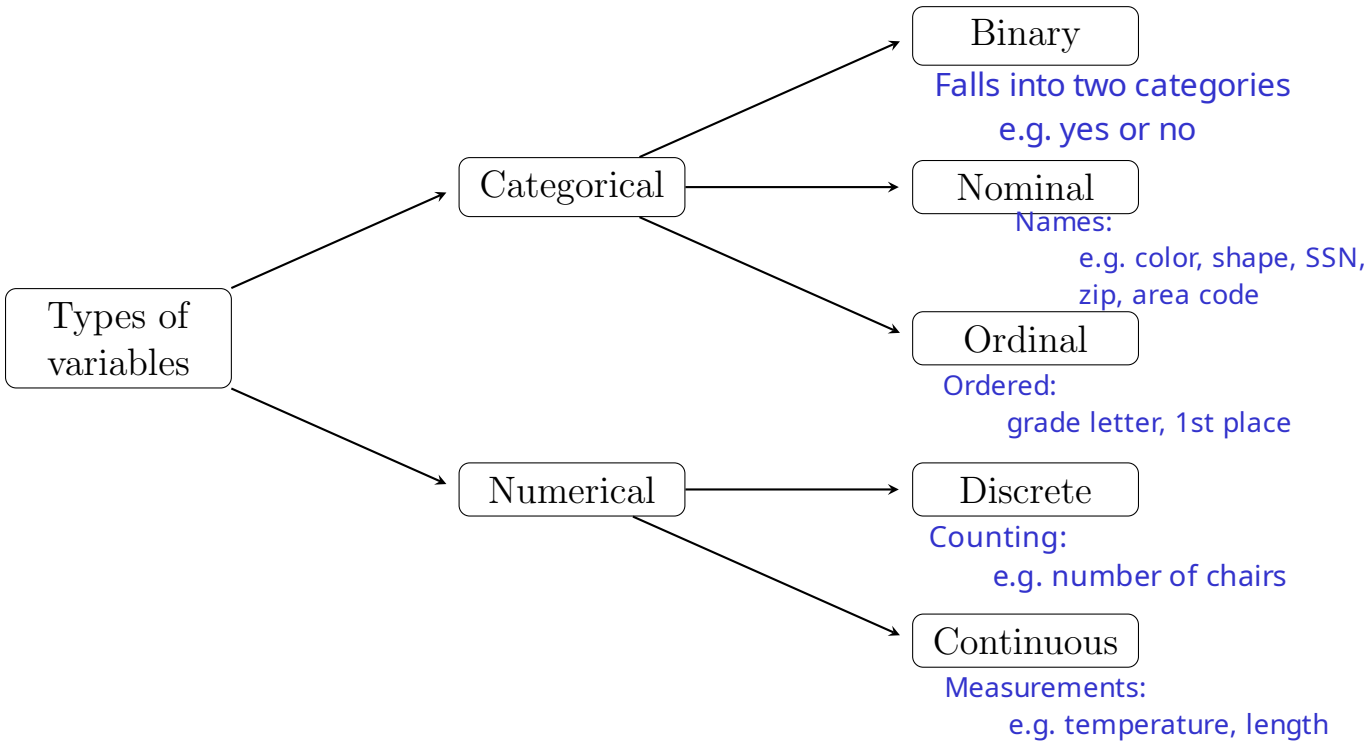
Coding categorical data using numbers:

Weight	Gender	Smoke
7.69	Female	No
0.88	Male	Yes
6.00	Female	No
7.19	Female	No
8.06	Female	No
7.94	Female	No

$\longrightarrow$

Weight	Female	Smoke
7.69	1	0
0.88	0	1
6.00	1	0
7.19	1	0
8.06	1	0
7.94	1	0

We can further break down variables into five types:



Example. Suppose a local store was interested in whether a new product would sell or not. The manager decided to take a random sample of 100 customers over a two-week period and asked each person whether they would buy the product or not and how many times would they buy the product over a six month period.

a) What is the population?

All possible customers

b) What is the sample?

100 customers

c) What are the variables?

would you buy, how often

d) Classify each variable as numerical or categorical.

would you buy
categorical (binary)

how often
numerical (discrete)

1.4: Organizing Categorical Data

Definition.
In the context of statistics, **frequency** is the number of times a value of a variable is observed in a data set.
Relative frequency (proportion) is a ratio of the frequency of a variable to the total frequency of the group desired. This can be left as a fraction, decimal, or percentage.

Example. The following **two-way table** contains the results of a national survey that asks American youths whether they wear a seat belt while driving or riding in a car:

	Male	Female	Total
Not Always	2	3	5
Always	4	8	10
Total	6	11	17

a) Find the total number of males, females, and total participants in this survey.

b) Identify the frequencies, and compute the percentages below:

number above

	Male	Female	Total
Not Always	11.8%	17.6%	29.4%
Always	23.5%	47.1%	70.6%
Total	35.3%	64.7%	100%

2/17=0.1176...

c) Are males or females more likely to take the risk of not wearing a seat belt?

Males: 3/6 → 33.%
Females: 4/11 → 27.3%

d) Should we use the frequencies or the relative frequencies to make comparisons?

Relative frequencies - makes data comparable

1.5: Collecting Data to Understand Causality

Definition.

- In an **observational study**, we observe individuals and measure variables of interest but do not attempt to influence the responses. (Observe but do not disturb)
- In a **controlled experiment**, we deliberately impose some treatment on (that is, do something to) individuals in order to observe their responses. Researchers assign subjects to a treatment group or control group.
- **Anecdotal evidence** is a story based on someone's experience.

- In an **observational study**, the researcher observes values of the response variable for the sampled subjects, without anything being done to the subjects (such as imposing a treatment).
- In short, an *observational study* merely observes rather than experiments with the study subjects.

Note: Anecdotal evidence and observational studies:

- NEVER point to causality (cause-and effect).
- Only point to an association between variables!

To establish cause-and-effect: Use a controlled experiment!

Definition.
Differences between two groups that could explain different experiences/outcomes are called **confounding variables** or **confounding factors**.

How to design a good experiment ("Gold standard" in experiments):

- Random allocation – participants randomly allocated to treatment and control group
- Use of a placebo if appropriate
  - A **placebo** is a fake treatment (e.g. sugar pill).
  - The **Placebo-Effect** is reacting to a treatment you haven't received.
- Blinding the study – used to avoid bias
  - Single blind – Researcher is unaware of treatment group
  - Double blind – Researcher and subjects are both unaware of treatment group
- Large sample size – accounts for variability